The BioGRID Interaction Database: 2008 update

Bobby-Joe Breitkreutz[1], Chris Stark[1], Teresa Reguly[1], Lorrie Boucher[1,2],
Ashton Breitkreutz[1], Michael Livstone[3], Rose Oughtred[3], Daniel H. Lackner[4],
Jürg Bähler[4], Valerie Wood[4], Kara Dolinski[3] and Mike Tyers[1,2,*]

[1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, M5G 1X5, [2]Department of Molecular Genetics, University of Toronto, Toronto, M5S 1A8, Canada, [3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA and [4]Wellcome Trust Sanger Institute, Cambridge CB10 1HH, UK

ABSTRACT

The Biological General Repository for Interaction Datasets (BioGRID) database (http://www.thebiogrid.org) was developed to house and distribute collections of protein and genetic interactions from major model organism species. BioGRID currently contains over 198 000 interactions from six different species, as derived from both high-throughput studies and conventional focused studies. Through comprehensive curation efforts, BioGRID now includes a virtually complete set of interactions reported to date in the primary literature for both the budding yeast _Saccharomyces cerevisiae_ and the fission yeast _Schizosaccharomyces pombe_. A number of new features have been added to the BioGRID including an improved user interface to display interactions based on different attributes, a mirror site and a dedicated interaction management system to coordinate curation across different locations. The BioGRID provides interaction data with monthly updates to Saccharomyces Genome Database, Flybase and Entrez Gene. Source code for the BioGRID and the linked Osprey network visualization system is now freely available without restriction.

INTRODUCTION

Protein interactions underlie cell structure, biochemical activity and dynamic behavior; in turn, myriad genetic interactions reflect the vast functional interconnectivities of the protein network (1). High-throughput technologies now generate large datasets of protein and genetic interactions, which compliment more conventional detailed investigations of cellular processes (2). The collation of various types of interaction data is essential for interrogation of system-level attributes (3), and to this end a number of important interaction databases have been developed (4–8). Previously, we described a database called 'Biological General Repository for Interaction Datasets' (BioGRID) (www.thebiogrid.org) to archive and distribute comprehensive collections of physical and genetic interactions (9).

The BioGRID has grown into a general resource for the research community with an average of 80 000 queries per month and millions of interactions downloaded per year. The 1 October 2007 version of BioGRID (v2.0.33) contains 198 791 (129 584 non-redundant) interaction records comprised of 137 834 (90 577 non-redundant) protein interactions and 60 957 (39 007 non-redundant) genetic interactions (Table 1). BioGRID provides full annotation support for 13 major model organism species (9), and currently houses interactions for _Saccharomyces cerevisiae_, _Schizosaccharomyces pombe_, _Caenorhabditis elegans_, _Drosophila melanogaster_, _Mus musculus_ and _Homo sapiens_. Gene annotation tables for all supported species are routinely updated to prevent ambiguous search results. Sources of data in BioGRID include publications that report high-throughput interaction datasets and many focused individual studies curated from inspection of the primary literature (2). Each interaction record in BioGRID contains experimental evidence codes and is linked to the supporting publication. In addition to the BioGRID website, all interactions in BioGRID are available through the dynamically linked Osprey visualization system, which can be used to query network organization in a user-defined fashion (10). BioGRID currently holds observer status

Table 1. Interactions in BioGRID

Species	Set	Nodes	Interactions (redundant)	Interactions (non-redundant)	Publication sources
S. cerevisiae	P	4936	64 799	38 298	4017
	G	3746	47 116	33 354	4339
	P + G	5196	111 915	71 652	7193
S. pombe	P	843	2631	1209	595
	G	689	2275	1769	733
	P + G	1057	4906	2978	1077
D. melanogaster	P	7091	22 834	22 479	140
	G	1199	10 689	3020	1515
	P + G	7548	33 523	25 499	1626
C. elegans	P	2816	4670	4557	8
	G	654	861	857	3
	P + G	3267	5531	5414	10
M. musculus	P	335	532	499	49
	G	3	2	2	2
	P + G	337	534	501	51
H. sapiens	P	8062	42 871	24 011	12 288
	G	29	20	11	17
	P + G	8069	42 891	24 022	12 296
All interactions	P	23 767	137 834	90 577	16 972
	G	6309	60 957	39 007	6599
	P + G	25 148	198 791	129 584	22 108

Redundant interactions include self edges, multiple sources and multiple experimental evidence types; non-redundant interactions exclude all duplicates. P, protein interactions; G, genetic interactions.

in the IMEx consortium of interaction databases (http://imex.sourceforge.net/).

LITERATURE CURATION

Comprehensive manual curation of the entire *S. cerevisiae* literature for protein and genetic interactions yielded 35 224 (21 281 non-redundant) protein interactions and 19 172 (13 963 non-redundant) genetic interactions (2), all of which are accessible through BioGRID. Comparison of the literature-curated protein interaction dataset to recent high-throughput studies (11–13) reveals a considerable degree of non-overlap, suggesting that many interactions remain to be validated and discovered in this yeast (Figure 1). We have since continued to curate the current *S. cerevisiae* literature and have added 29 575 (17 017 non-redundant) protein and 27 994 (19 391 non-redundant) genetic interactions to BioGRID since the original curation effort. Updates to BioGRID are made on the first of every month; each release of BioGRID is date stamped and archived for comparative purposes. All *S. cerevisiae* interactions deposited in BioGRID are immediately imported into the Saccharomyces Genome Database (SGD) with associated citations and evidence codes (14). Additional interaction attributes, including post-translational modifications associated with protein interactions and specific phenotypes associated with each genetic interaction, are currently being annotated and will be released for the entire *S. cerevisiae* dataset in the near future.

To complement the comprehensive *S. cerevisiae* dataset, we have recently completed exhaustive manual curation of the *S. pombe* literature. Interactions were classified based on the same experimental evidence codes for protein and genetic interactions used previously (2). This effort yielded 2631 (1209 non-redundant) protein interactions and 2275 (1769 non-redundant) genetic interactions, as derived from 1077 publications. This new dataset has recently been deposited in BioGRID and, as for *S. cerevisiae*, will be updated on a monthly basis and provided to the *S. pombe* genome database (GeneDB) currently hosted by the Sanger Institute at www.genedb.org/genedb/pombe/ (15). Comparison of orthologous interactions between these evolutionary distant yeasts should prove informative for biological network structure and function. Imminent high-throughput studies in *S. pombe* should rapidly elaborate the cellular interaction network in this organism (16–18).

In addition to systematic yeast curation, we have also undertaken partial interaction curation for higher species, for example, *D. melanogaster* and *H. sapiens* (Table 1). These curation efforts are often focused on specific aspects of biology and are in part guided by Gene Ontology inference codes for protein and genetic interactions (19) and the Textpresso text-mining algorithm (20). Other species interactions are added to BioGRID on an ongoing basis and, when available, released in monthly BioGRID updates. Contributions of curated interaction datasets from any species for deposition in the BioGRID are welcomed (www.thebiogrid.org).

DATABASE IMPROVEMENTS

We have expanded accessibility to interaction data in BioGRID via a primary mirror site at the SGD colony in Princeton (http://grid.princeton.edu/). In addition, source code for BioGRID and Osprey has been made available without restriction at SourceForge. BioGRID data files are currently linked to SGD, Flybase and NCBI, to which we provide automatic monthly updates.
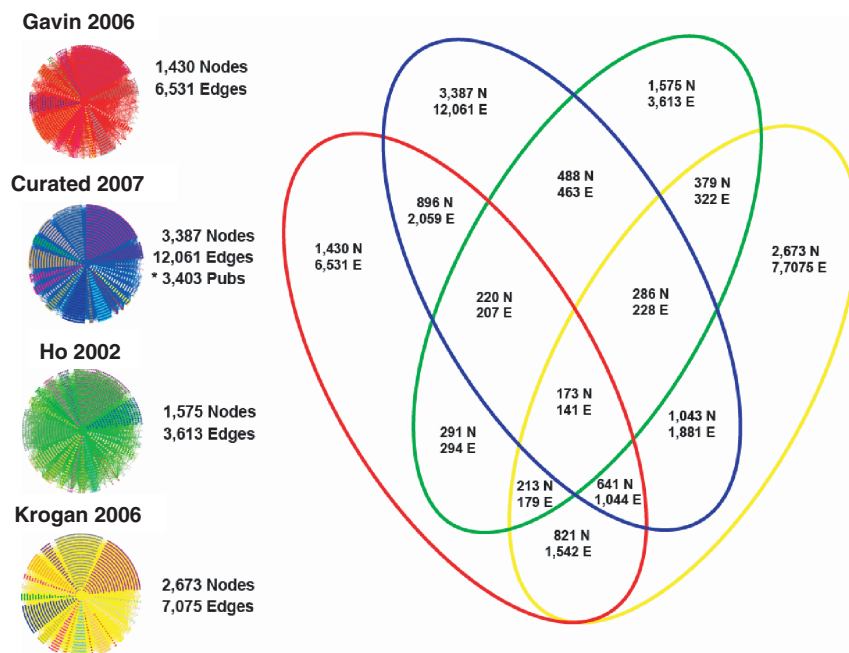
Figure 1. Overlap between various high-throughput protein interaction datasets (11–13) and the literature-curated interaction dataset as of 1 January 2007 (2). Dataset manipulation was performed with the Osprey visualization system (10). All datasets are available for download from www.thebiogrid.org. N, interaction node and E, interaction edge.

Analogous relationships are underway with the *Arabidopsis* Information Resource (TAIR) and *S. pombe* GeneDB (15,21). We will endeavor to fulfill all requests for custom datasets for export to other model organism databases; the download page at BioGRID contains examples of existing datasets created for export.

The dataset download system for the BioGRID is now powered by Asynchronous JavaScript and XML (AJAX), such that downloads are available for every search result page and publication in the database. Users may thus download interactions associated with any gene or all interactions reported by a single publication. Further data subsets organized by experimental system or by organism, including a recently described multivalidated yeast–protein interaction dataset (22), are also available on the BioGRID download page.

The tabular user interface of BioGRID has been improved through implementation of AJAX techniques. The interface now provides the option to narrow search results to quantitative datasets; this feature will soon be elaborated to enable user-defined search criteria according to data type, evidence codes and data source. The ability to expand hidden fields with a single mouse click to provide greater detail, such as for Gene Ontology classifications (23), has also been added. Search results now include bait and hit designations to indicate the directionality of interactions. Additional annotation features including phenotype, post-translational modification, domains and motifs are currently under construction.

THE INTERACTION MANAGEMENT SYSTEM

We have implemented an interaction management system (IMS) to support multiple simultaneous curators for each species supported by BioGRID. The IMS is a multiuser web-based application written in PHP that interfaces directly with the BioGRID. An intuitive graphical interface allows curators to quickly record interactions from an automatically updated list of publications. All interactions added via the IMS are verified against current annotation tables to eliminate errors and ambiguity in gene nomenclature. The IMS instantly commits new interactions to the BioGRID update pipeline, unless specified otherwise; interactions are collated each month and released as updates to the primary BioGRID and mirror sites, as well as model organism collaboration sites. Interactions may also be removed or modified in each monthly build, for example, in response to community feedback. All retired datasets are archived on the BioGRID downloads page in case the need for back-comparison arises.

FUTURE DEVELOPMENT

We will continue to curate interactions from major model organism species, with a view to comprehensive back-curation, as we have done for *S. cerevisiae* and *S. pombe*. Further refinement of tools and display features in the BioGRID graphical user interface based on a flexible record tag structure will enable greater control over data views and downloadable datasets by the user. New plugins for data visualization are under development for Osprey, Cytoscape (24) and the Edinburgh Pathway Editor (25), which will allow more sophisticated interrogation of interaction networks. In order to facilitate dissemination of our open source software tools, we will strive for compatibility with the Generic Model Organism Database (GMOD) project (26). Finally, we will continue to develop

our record structures in compliance with the Proteomics Standards Initiative Molecular Interactions (PSI-MI) standard (27,28).

REFERENCES

1. Jorgensen,P., Breitkreutz,B.J., Breitkreutz,K., Stark,C., Liu,G., Cook,M., Sharom,J., Nishikawa,J.L., Ketela,T. *et al.* (2003) Harvesting the genome's bounty: integrative genomics. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 431–443.
2. Reguly,T., Breitkreutz,A., Boucher,L., Breitkreutz,B.J., Hon,G.C., Myers,C.L., Parsons,A., Friesen,H., Oughtred,R. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.*, **5**, 11.
3. Batada,N.N., Reguly,T., Breitkreutz,A., Boucher,L., Breitkreutz,B.J., Hurst,L.D. and Tyers,M. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.*, **4**, e317.
4. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
5. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
6. Mewes,H.W., Frishman,D., Mayer,K.F., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
7. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
8. Chatr-Aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
9. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
10. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol.*, **4**, R22.
11. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
12. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
13. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.
14. Nash,R., Weng,S., Hitz,B., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.*, **35**, D468–D471.
15. Wixon,J. and Wood,V. (2006) Tools and resources for Sz. pombe: a report from the 2006 European Fission Yeast Meeting. *Yeast*, **23**, 901–903.
16. Gould,K.L., Ren,L., Feoktistova,A.S., Jennings,J.L. and Link,A.J. (2004) Tandem affinity purification and identification of protein complex components. *Methods*, **33**, 239–244.
17. Matsuyama,A., Arai,R., Yashiroda,Y., Shirai,A., Kamata,A., Sekido,S., Kobayashi,Y., Hashimoto,A., Hamamoto,M. *et al.* (2006) ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat. Biotechnol.*, **24**, 841–847.
18. Roguev,A., Wiren,M., Weissman,J.S. and Krogan,N.J. (2007) High-throughput genetic interaction mapping in the fission yeast *Schizosaccharomyces pombe*. *Nature Methods*, **4**, 861–866.
19. Drabkin,H.J., Hollenbeck,C., Hill,D.P. and Blake,J.A. (2005) Ontological visualization of protein–protein interactions. *BMC Bioinformatics*, **6**, 29.
20. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
21. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
22. Batada,N.N., Reguly,T., Breitkreutz,A., Boucher,L., Breitkreutz,B.J., Hurst,L.D. and Tyers,M. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biol.*, **5**, e154.
23. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
24. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
25. Sorokin,A., Paliy,K., Selkov,A., Demin,O.V., Dronov,S., Ghazal,P. and Goryanin,I. (2006) The Pathway Editor: a tool for managing complex biological networks. *IBM J. Res. Dev.*, **6**, 561–575.
26. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
27. Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The proteomics standards initiative. *Proteomics*, **3**, 1374–1376.
28. Orchard,S., Salwinski,L., Kerrien,S., Montecchi-Palazzi,L., Oesterheld,M., Stumpflen,V., Ceol,A., Chatr-Aryamontri,A., Armstrong,J. *et al.* (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **25**, 894–898.